Astin Bulletin (2024), **54**, pp. 263–279
doi:10.1017/asb.2024.6

International Actuarial Association
Association Actuarielle Internationale

RESEARCH ARTICLE

Integration of traditional and telematics data for efficient insurance claims prediction

Hashan Peiris[1], Himchan Jeong[1], Jae-Kwang Kim[2] and Hangsuck Lee[3]

[1]Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, British Columbia, V5A1S6, Canada, [2]Department of Statistics, Iowa State University, Ames, Iowa, 50011, USA and [3]Department of Mathematics/Actuarial Science, Sungkyunkwan University, Seoul, 03063, South Korea
Corresponding author: Himchan Jeong; Email: himchan_jeong@sfu.ca

Received: 19 April 2023; **Revised:** 14 December 2023; **Accepted:** 19 January 2024; **First published online:** 15 February 2024

Keywords Favorable selection; automobile insurance; data integration; driver telematics; missing data analysis

Abstract

While driver telematics has gained attention for risk classification in auto insurance, scarcity of observations with telematics features has been problematic, which could be owing to either privacy concerns or favorable selection compared to the data points with traditional features.

To handle this issue, we apply a data integration technique based on calibration weights for usage-based insurance with multiple sources of data. It is shown that the proposed framework can efficiently integrate traditional data and telematics data and can also deal with possible favorable selection issues related to telematics data availability. Our findings are supported by a simulation study and empirical analysis in a synthetic telematics dataset.

1. Introduction

Telematics generates data related to many variables characterized for each driver, including total miles driven, the number of sudden brakes or accelerations, and at what time they are driving. With technological advancements in the automobile industry with driver telematics, the insurance industry can add new features to the databases along with the traditional features that will be used in claim predictions and risk classifications in a unified frame. In this regard, it is required to consider a framework to deal with multiple data sources that contain traditional and/or telematics features for insurance ratemaking, which is one of the main contributions of this paper.

The usage-based insurance (UBI) is an innovative product in the insurance industry based on technological advances to assess the risk profile of a driver. Past studies elaborate on the additional value of telematics-derived information to provide improved claims predictions, risk classification, and premium assessments. Ayuso *et al.* (2014) compare driving behaviors of novice and experienced young drivers with pay-as-you-drive policies using few telematics variables as well as traditional variables. Furthermore, Ayuso *et al.* (2016) examine gender discrimination in the risk of accidents using the same dataset. Baecke and Bocca (2017) illustrate the use of telematics variables to decide the risk premium and state that at least three months of data are enough to obtain efficient risk estimates. Verbelen *et al.* (2018) depict the importance of telematics variables, which are based on driving habits, in predicting the frequency of claims. Gao *et al.* (2019) show the predictive power of telematics covariates extracted from speed-acceleration heat maps in the modeling of claim frequency and support the use of telematics features for insurance pricing.

Moreover, insurance companies have access to large datasets related to policyholders that contain traditional characteristics, as driver demographics and vehicle characteristics. However, a telematics

dataset can have fewer data points than a traditional dataset, as the number of telematics related policyholders is low. Guillen *et al.* (2021) use a modeling approach for insurance ratemaking using traditional and telematics data but is limited to a small number of features, as available data are limited. Ma *et al.* (2018) mention that the lack of availability of telematics data is a challenge in identifying the factors of policyholder behavior in ratemaking. While providing a compact description of the insurability of risk using telematics data, Eling and Kraft (2020) highlight some actions that can increase the number of telematics-based policyholders. Hence, there is a scarcity of telematics data when compared to traditional data.

In this regard, it is natural to expect that insurers need to deal with two types of datasets: traditional datasets with fewer features and more observations from non-UBI insureds and telematics datasets with more features and fewer observations from UBI insureds. One could argue that insurers could potentially treat UBI and non-UBI insureds as separate groups and it suffices to analyze two types of datasets separately, as more and more people with low risk would move to UBI over time and form a natural market segmentation due to asymmetric information (Rothschild and Stiglitz, 1978). According to Holzapfel *et al*., 2023), however, the market share of UBI contracts remains relatively low and stands around at about 5%, whereas UBI contracts have been accessible to policyholders for over twenty years (NAIC, 2015; MarketsandMarkets, 2021). At the very least, the situation where there are far fewer UBI subscribers than non-subscribers can last longer than expected, and therefore the data integration framework that we propose could be valid for a considerable period of time in the future. Further, it is natural to expect that a policyholder may move back and forth between a UBI and a non-UBI contract upon renewal (Śliwiński and Kuryłowicz, 2021). Therefore, it is worthwhile to investigate the available datasets jointly to better understand the characteristics of the population, compared to a separate analysis of the traditional and telematics datasets that implicitly assume time-invariant business mix between UBI and non-UBI contracts of an auto insurance company.

Data integration techniques enable combining information from a few data sources into one. According to Yang and Kim (2020), it leads to the incorporation of information from different samples to achieve efficiency in estimations under finite population inference while handling potential selection biases. And Husnjak *et al.* (2015) recognize that the integration of telematics data with traditional data can help to realize the full potential of telematics data. Thus, Ayuso *et al.* (2019) and Gao *et al.* (2022) propose two-step approaches that use telematic characteristics to improve a regression model that only incorporates traditional ratemaking factors.

Although these approaches are straightforward and readily available, they might be problematic when the availability of telematics features depends on the riskiness of the policyholders due to possible favorable selection. For example, Denuit *et al.* (2019) state that low-risk drivers would favor telematics insurance products. And Duval *et al.* (2023) mention that the attraction of safer drivers is beneficial for the insurer as it could lower the claim cost. However, this situation may result in missing some insights about more risky drivers in terms of an analytical point of view. According to Cohen and Siegelman (2010), one can expect that the information asymmetry between insurers and policyholders may lead to favorable selection in the sampling mechanism of observations with telematics features as less risky drivers are more likely to provide telematics data for possible premium discounts.

Indeed, consideration and collection of telematics data are relatively recent, and there are still ongoing concerns about privacy issues, which make many policyholders reluctant to agree on the provision of their telematics data to insurers. In this regard, Dewri *et al.* (2013) state privacy concerns that can arise when using telematics data for driving habits. Also Duri *et al.* (2002) mention that there is a tendency to observe a decrease in the amount of telematics data due to privacy concerns, which is a similar trend among web users with privacy concerns. In a similar way, Milanović *et al.* (2020) imply that policyholders who are willing to provide telematics data tend to have less concern about privacy issues. Thus, we can observe a selection bias in the telematics dataset due to privacy issues as well as the favorable selection.

https://doi.org/10.1017/asb.2024.6 Published online by Cambridge University Press

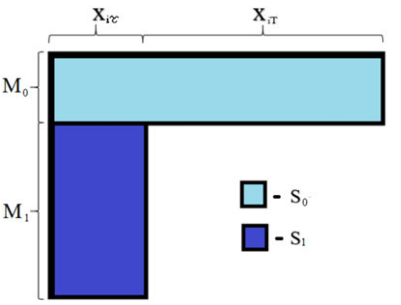

Figure 1. *Pictorial visualization of* $\mathcal{S}_0$, $\mathcal{S}_1$, $\boldsymbol{x}_{i\tau}$, *and* $\boldsymbol{x}_{iT}$.

In summary, the following objectives of the proposed research are recognized. First, we propose a framework based on the estimation of the propensity score to combine information from multiple datasets in insurance ratemaking considering the scarcity of telematics data and possible favorable selection regarding the availability of telematics data. Then we introduce an algorithm to integrate a traditional insurance claim dataset and a telematics dataset based on a calibration equation approach in detail. Finally, we test the validity and applicability of the proposed framework through a simulation study and empirical analysis of a synthetic telematics dataset. Consequently, we hope that the proposed method can help insurance companies effectively use multiple sources of data for better risk classification.

The rest of this article is organized as follows. Section 2 provides a detailed description of the problem and the corresponding data structure with the missing mechanism. In Section 3, the proposed framework for data integration is developed based on a calibration equation approach with information projection to model the claim count data. Section 4 provides a simulation study to assess the effects of the proposed approach compared to four preexisting approaches. Section 5 conducts an empirical analysis with a synthetic telematics data portfolio that is emulated from real data, to assess the applicability of the proposed approach in practice. Section 6 concludes the paper with some constructive remarks.

2. Data structure and problem description

This study focuses on two data sources as discussed in Section 1. $\mathcal{S}_0$, a small dataset with M_0 observations that contains both telematics and traditional features. And $\mathcal{S}_1$, a large dataset with M_1 number of observations that contains only traditional features. We also assume that the finite population $\mathcal{S}$ consists of both $\mathcal{S}_0$ and $\mathcal{S}_1$ and that the total number of observations in $\mathcal{S}$ is $M = M_0 + M_1$.

We denote traditional features of a policyholder i as $\mathbf{x}_{i\tau}$, (available both in $\mathcal{S}_0$ and $\mathcal{S}_1$) and telematics features of a policyholder i as $\mathbf{x}_{iT}$, which is only available in $\mathcal{S}_0$. Using these features, all the corresponding features of the study can be denoted as a vector, $\mathbf{x}_i = (\mathbf{x}_{i\tau}, \mathbf{x}_{iT})$. A summary of the description of the data is given in Figure 1.

Note that the observability of $\mathbf{x}_{iT}$ could depend on the risk profile of a policyholder i, which could make the sampling mechanism of $\mathcal{S}_0$ from the population subject to selection biases. As mentioned in the previous section, there have been possible concerns about providing telematics records, such as privacy and security issues; hence, it is natural to expect that a policyholder might not be willing to provide their telematics records to the insurer unless the expected benefits from the provision outweigh the possible concerns. Therefore, we can think of the following conjectures:

- Those who are younger tend to agree to provide their telematics records more, as they could be less reluctant to technology or the compensation for disclosing privacy to get a UBI policy is lower according to Derikx *et al.* (2016). It implies that the probability of observing a data point in $\mathcal{S}_0$ could be negatively correlated with the driver's age.

https://doi.org/10.1017/asb.2024.6 Published online by Cambridge University Press

- Those who are less risky[1] tend to agree to provide their telematics records or the UBI policyholders tend to be less risky drivers according to Reimers and Shiller (2019) and Cather (2020), so that the accessibility of $\mathbf{x}_T$ is prone to favorable selection. It implies that the probability of observing a data point in $\mathcal{S}_0$ could be negatively correlated with the number of claims (n_i).
- Those who drive less frequently tend to agree to provide their telematics records more since the premium is low in UBI products as in Boucher *et al.* (2013). It implies that the probability of observing a data point in S_0 could be negatively correlated with the self-perceived mileage.

While our main task is neither to detect possible selection biases in the availability of telematics features nor prove such conjectures, we consider the situations where such conjectures could hold and discuss the benefits of the proposed framework compared to preexisting benchmarks in various situations.

3. Methodology

The general framework that we follow to estimate the model parameters using the proposed method is briefly described in this section. We are interested in estimating $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ in the regression model $E(N_i \mid \mathbf{x}_i) = m(\mathbf{x}_i\boldsymbol{\beta}) = m(\mathbf{x}_{i\tau}\boldsymbol{\beta}_1 + \mathbf{x}_{iT}\boldsymbol{\beta}_2)$, where $m(\cdot)$ is a known function and $\boldsymbol{\beta}$ is an unknown parameter while N_i is the observed number of claims for a policyholder i with $i = 1, \ldots, M$. We assume that N_i are independently distributed with a Poisson distribution with mean μ_i.[2] Using the canonical link function as it is given in Agresti (2003), we can express $m(\cdot) = \exp(\cdot)$. Let t_i be an exposure variable associated with i^{th} claim count like the duration of a policy. Then, η_i is the average number of claims per the ith duration. Now we can redefine the regression model in terms of η_i as

$$\log(\eta_i) = \sum_j \beta_j x_{ij} = \mathbf{x}_i\boldsymbol{\beta},$$

where x_{ij} is the jth feature of the policyholder i and $x_{i0} = 1$. Thus, we can reform this model using the definition $\mu_i = t_i\eta_i$[3] as

$$\log(\mu_i) = \log(t_i) + \sum_j \beta_j x_{ij} = \log(t_i) + \mathbf{x}_i\boldsymbol{\beta}. \tag{3.1}$$

Now, using model (3.1), the census estimating equation for $\boldsymbol{\beta}$ can be written as

$$\sum_{i=1}^{M} U(\boldsymbol{\beta}; \mathbf{x}_i, n_i) = 0, \tag{3.2}$$

where $U(\boldsymbol{\beta}; \mathbf{x}, n) = \{n - t\exp(\mathbf{x}\boldsymbol{\beta})\}\mathbf{x}$ is the estimating function for $\boldsymbol{\beta}$ with a Poisson distribution. However, as mentioned in Section 2, $\mathbf{x}_{iT}$ (which corresponds to the telematics features of a policyholder i) is subject to missingness and only observable in $\mathcal{S}_0$. In this regard, one can consider the following equation to estimate $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ simultaneously:

$$\sum_{i\in\mathcal{S}_0} \omega_i U(\boldsymbol{\beta}; \mathbf{x}_i, n_i) = 0, \tag{3.3}$$

where ω_i is a propensity weight to handle possible selection biases.

[1]Note that riskier drivers can be attracted to UBI policies provided there is a upfront discount for choosing a UBI contract, and they can continue to adjust their driving behaviors to maintain such an incentive, as mentioned in Duval *et al.* (2023). However, we do not assume presence of a upfront discount for choosing a UBI policy in our article.

[2]While the proposed method is independent of the distribution of N_i as it is based on general calibration and estimating equations, here we use a Poisson distribution assumption to focus on the impact of the proposed data integration approach.

[3]Note that $E(N_i \mid \mathbf{x}_i) = \mu_i$.

https://doi.org/10.1017/asb.2024.6 Published online by Cambridge University Press

To incorporate the partial information in $\mathcal{S}_1$ where we only observe $\mathbf{x}_{i\tau}$ and n_i, we wish to construct the propensity weight $\omega_i = \omega(\mathbf{x}_{i\tau}, n_i)$ in $\mathcal{S}_0$ such that

$$\sum_{i\in\mathcal{S}_0} \omega_i U(\boldsymbol{\beta}; \mathbf{x}_i, n_i) = \sum_{i=1}^{M} \left[\delta_i U(\boldsymbol{\beta}; \mathbf{x}_i, n_i) + (1-\delta_i)\bar{U}(\boldsymbol{\beta}; \mathbf{x}_{i\tau}, n_i)\right], \tag{3.4}$$

where $\delta_i = \mathbb{I}(i \in \mathcal{S}_0)$ and $\bar{U}(\boldsymbol{\beta}; \mathbf{x}_{i\tau}, n_i) = E\{U(\boldsymbol{\beta}; \mathbf{x}_i, n_i) \mid \mathbf{x}_{i\tau}, n_i\}$. The propensity score (PS) is defined as $\omega_i = 1/Pr(\delta_i = 1|\mathbf{x}_i, n_i)$. The property of the propensity score estimating equation in (3.4) is called self-efficiency, as it leads to an efficient estimation of $\boldsymbol{\beta}$ as long as the conditional expectation in $E\{U(\boldsymbol{\beta}; \mathbf{x}_i, n_i) \mid \mathbf{x}_{i\tau}, n_i\}$ is correct.

Here, we assume that the sampling mechanism for $\mathcal{S}_0$ is missing at random (MAR) in the sense of Rubin (1976). That is, we assume

$$\delta \perp \mathbf{x}_T \mid (n, \mathbf{x}_\tau).$$

To find ω_i satisfying (3.4), we first find the basis functions satisfying

$$E\{U(\boldsymbol{\beta}; \mathbf{x}_i, n_i) \mid \mathbf{x}_{i\tau}, n_i\} \in \text{span}\{b_1(\mathbf{x}_{i\tau}, n_i), \ldots, b_L(\mathbf{x}_{i\tau}, n_i)\}, \tag{3.5}$$

where the span implies that the conditional expectation is represented by a combination of basis functions, b_l, that are formed only using the traditional features and observed number of claims where $l = 1, \ldots, L$.[4] Under (3.5), estimating the conditional expectation $E\{U(\boldsymbol{\beta}; \mathbf{x}_i, n_i) \mid \mathbf{x}_{i\tau}, n_i\}$ is somewhat tricky as $U(\boldsymbol{\beta}; \mathbf{x}_i, n_i)$ involves unknown parameter $\boldsymbol{\beta}$. To avoid this difficulty, we consider an alternative method using (3.4) without estimating the conditional expectation.

To achieve this goal, using the basis functions in (3.5), we impose the following system of equations

$$\sum_{i\in\mathcal{S}_0} \omega_i [1, b_{1i}, \cdots, b_{Li}] = \sum_{i=1}^{M} [1, b_{1i}, \cdots, b_{Li}], \tag{3.6}$$

as a constraint for finding the propensity weights ω_i in (3.4), where $b_{li} = b_l(\mathbf{x}_{i\tau}, n_i)$ is a vector of integrable functions of traditional features and $[1, b_{1i}, \cdots, b_{Li}]$ is a $L+1$ dimensional vector. To be specific, we take $[1, b_{1i}, \cdots, b_{Li}] = [x_{i\tau}, n_i \cdot x_{i\tau}]$ inspired by the form of Poisson score function, which implies $L = 2v+1$ where v is the number of features in $x_{i\tau}$. Constraint (3.6) is often called the covariate-balancing property (Imai and Ratkovic, 2014) in the context of causal inference, which enables an efficient estimation of the propensity score by assuring that the distributions of available covariates in propensity weighted sample and the population are similar. The following proposition shows that the covariate balancing is a sufficient condition for self-efficiency in (3.4).

Proposition 1. *Suppose that the estimating function satisfies (3.5). Then, any weights satisfying (3.6) satisfies the self-efficiency in (3.4).*

Proof. See Appendix A. □

Now, to uniquely determine ω_i, we can use the information projection of Wang and Kim (2021) under the constraint (3.6) to get

$$\omega_i = 1 + \frac{M_1}{M_0} \exp\{\phi_0 + \phi_1 b_{1i} + \cdots + \phi_L b_{Li}\}, \tag{3.7}$$

where $M_0 = \sum_{i=1}^{M} \delta_i$, $M_1 = M - M_0$ and $\boldsymbol{\phi} = (\phi_0, \cdots, \phi_L)$ is an unknown parameter. The parameters are estimated by solving the calibration equation in (3.6).

[4]Note that both determination of optimal L and construction of b_l are still open questions as larger values of L would imply richer information to span the space of conditional expectations at the expense of estimation efficiency. That is, increasing L can reduce the chance of misspecification bias but may increase variance due to parameter estimation.

https://doi.org/10.1017/asb.2024.6 Published online by Cambridge University Press

Once $\phi_0, \cdots, \phi_L$ are estimated by (3.6) and (3.7), we can use

$$\hat{\omega}_i = 1 + \frac{M_1}{M_0} \exp\left\{\hat{\phi}_0 + \hat{\phi}_1 b_{1i} + \cdots + \hat{\phi}_L b_{Li}\right\}$$

as the final propensity weights for estimating $\boldsymbol{\beta}$ using (3.8):

$$\sum_{i \in \mathcal{S}_0} \hat{\omega}_i(\boldsymbol{\phi}) U(\boldsymbol{\beta}; \mathbf{x}_i, n_i) = 0. \tag{3.8}$$

Because the propensity weights satisfy the calibration equation in (3.6), it satisfies the self-efficiency without estimating the regression coefficients $\hat{\boldsymbol{\alpha}}$[5] in the working model

$$E\{U(\boldsymbol{\beta}; \mathbf{x}_i, n_i) \mid \mathbf{x}_{i\tau}, n_i\} = \alpha_0 + \sum_{l=1}^{L} \alpha_l b_l(\mathbf{x}_{i\tau}, n_i).$$

But, the vector space spanned in (3.5) implicitly assumes a regression model that is

$$U_i = \alpha_0 + \sum_{l=1}^{L} \alpha_l b_{li} + e_i,$$

for some $\alpha_0, \alpha_1, \cdots, \alpha_L$, where $U_i = U(\beta; \mathbf{x}_i, n_i)$ and e_i is the error term satisfying $E(e_i) = 0$. Since $U_i = (U_{i1}, \cdots, U_{ip})'$, the above model changes to

$$U_{ij} = \alpha_{0j} + \sum_{l=1}^{L} \alpha_{lj} b_{li} + e_{ij}$$

where $e_{ij} \sim (0, V_j)$.

Then, $\hat{U}_i = \hat{\alpha}_0 + \sum_{l=1}^{L} \hat{\alpha}_l b_{li}$ and $\hat{\alpha}_l$ $(l = 0, 1, \cdots, L)$ are chosen to minimize

$$\sum_{i \in S_0} g_i(\hat{\phi}) \left\{ U_i - \alpha_0 - \sum_{l=1}^{L} \alpha_l b_l(\mathbf{x}_{i\tau}, n_i) \right\}^2$$

with respect to $(\alpha_0, \alpha_1, \cdots, \alpha_L)$, where $g_i(\hat{\phi}) = \exp\left\{\hat{\phi}_0 + \hat{\phi}_1 b_{1i} + \cdots + \hat{\phi}_L b_{Li}\right\}$. Thus, $\hat{U}_i$ satisfies

$$\sum_{i \in S_0} \left(U_i - \hat{U}_i\right) g_i(\hat{\phi}) = 0. \tag{3.9}$$

Proposition 2. *The proposed weight in (3.7) satisfies self-efficiency in (3.4) when $\bar{U}_i$ is replaced with $\hat{U}_i$.*

Proof. See Appendix B. □

Now, to improve this proposed method, we may use the information of model variance. Suppose that $V(e_i) = v_i$ is available, then we can use

$$\hat{\omega}_i = 1 + \frac{M_1}{M_0} \exp\left\{\hat{\phi}_0 + \hat{\phi}_1 b_{1i} + \cdots + \hat{\phi}_L b_{Li}\right\} \frac{1}{v_i} \tag{3.10}$$

as the final propensity weights for estimating θ. It can still achieve (3.4) where $\hat{\alpha}_l (l = 0, 1, \cdots, L)$ minimizes

$$\sum_{i \in S_0} g_i(\hat{\phi}) \left\{ U_i - \alpha_0 - \sum_{l=1}^{L} \alpha_l b_l(\mathbf{x}_{i\tau}, n_i) \right\}^2 \frac{1}{v_i}$$

[5] $\hat{\boldsymbol{\alpha}}$ is the estimated regression coefficients of the (linear) working model explained in Equation (3.5).

https://doi.org/10.1017/asb.2024.6 Published online by Cambridge University Press

We can simply use the class in (3.10) as a class of calibration weights and choose $v_i = f(\mathbf{b}_i)$ such that (3.6) holds and reduces the variance (by downweighting the large weights). One way is to use v_i from the conditional variance of U_i given the covariates.

Now, the estimation scheme for the study is listed in order according to the requirements of the estimation process at each step.

1. Find $\mathcal{H} = \text{span}\{b_1(\mathbf{x}_{i\tau}, n_i), \ldots, b_L(\mathbf{x}_{i\tau}, n_i)\}$ such that $E\{U(\boldsymbol{\beta}; \mathbf{x}_i, n_i) \mid \mathbf{x}_{i\tau}, n_i\} \in \mathcal{H}$, where $U(\boldsymbol{\beta}; \mathbf{x}_i, n_i)$ is the estimating function for $\boldsymbol{\beta}$.
2. Find v_i using a suitable method.
3. Obtain $\hat{\boldsymbol{\phi}}$ by solving
$$\sum_{i \in \mathcal{S}_0} \left\{ 1 + \frac{M_1}{M_0} \exp(\phi_0 + \phi_1 b_{1i} + \cdots + \phi_L b_{Li}) \frac{1}{v_i} \right\} [1, b_{1i}, \cdots, b_{Li}] = \sum_{i=1}^{M} [1, b_{1i}, \cdots, b_{Li}],$$
4. Obtain $\hat{\boldsymbol{\beta}}$ by solving
$$\sum_{i \in \mathcal{S}_0} \hat{\omega}_i(\hat{\phi}) U(\boldsymbol{\beta}; \mathbf{x}_i, n_i) = 0$$
where $\hat{\omega}_i(\hat{\phi}) = 1 + \frac{M_1}{M_0} \exp\left\{ \hat{\phi}_0 + \hat{\phi}_1 b_{1i} + \cdots + \hat{\phi}_L b_{Li} \right\} \frac{1}{v_i}$.

The estimation of the standard error of $\hat{\boldsymbol{\beta}}$ is presented in Appendix C.

4. Simulation study

In this section, we use a hypothetical and less complex finite population to test the validity and applicability of the proposed method. More specifically, it allows us to quantify the estimation performance of regression coefficients with the proposed model (compared to the benchmarks) using finite samples from a predetermined distribution. In this regard, we assume three hypothetical scenarios in which traditional features are fully available while telematics features are partially available, depending on the sampling mechanism of observations with telematics information. We generate a finite population of size 100,000 with the following specification:

$$\begin{aligned}
&N_i \sim \mathcal{P}(\mu_i), \quad \log \mu_i = \mathbf{x}_{i\tau} \boldsymbol{\beta}_1 + \mathbf{x}_{iT} \boldsymbol{\beta}_2, \\
&\boldsymbol{\beta}_1 = (\beta_0, \beta_{A1}, \beta_{A2}, \beta_G, \beta_M), \quad \boldsymbol{\beta}_2 = \beta_T, \\
&\mathbf{x}_{i\tau} = (1, x_{Ai}, x_{Ai}^2, x_{Gi}, x_{Mi}), \quad \mathbf{x}_{iT} = x_{Ti}, \\
&x_{Ai} \sim \mathcal{U}(0.18, 0.81),\ x_{Gi} \sim \mathcal{B}er(0.6),\ x_{Mi} \sim \mathcal{G}(2, 1),\ x_{Ti} \sim \mathcal{N}(0, 1), \\
&\beta_0 = -1.3,\ \beta_{A1} = -4,\ \beta_{A2} = 3.4,\ \beta_G = 0.1,\ \beta_M = 0.1,\ \beta_T = 0.5,
\end{aligned}$$

where $\mathcal{P}$, $\mathcal{U}$, $\mathcal{B}er$, $\mathcal{N}$, and $\mathcal{G}$ refer to Poisson, uniform, Bernoulli, normal, and gamma[6] distributions, respectively. Here, x_{Ai} refers to a traditional continuous variable with quadratic effect (e.g., driver's age), x_{Gi} refers to a traditional binary variable (e.g., gender), x_{Mi} refers to a traditional variable like self-perceived mileage, and x_{Ti} refers to a telematics variable that has significant impacts on the risk profile. Let $\mathcal{S}^*$ be the finite population generated according to the notation used in Section 2. Once a finite population is generated, the following scheme is applied to split the data.

1. First, 10% of the data points are set aside where $\{N_i, \mathbf{x}_{i\tau}, \mathbf{x}_{iT}\}$ are all available, which is equivalent to $\mathcal{S}_0$ in Section 2. Depending on the assumption of availability of telematics information, we apply the following four sampling schemes of $\mathcal{S}_0$:

[6]Here we used the parameterization of Klugman *et al.* (2012) such that $\mathbb{E}[Z] = \alpha\theta$ and $Var[Z] = \alpha\theta^2$ if $Z \sim \mathcal{G}(\alpha, \theta)$.

https://doi.org/10.1017/asb.2024.6 Published online by Cambridge University Press

- **Random selection**: Data points assigned to $\mathcal{S}_0$ are chosen at random,
- **Age selection**: Each data point assigned to $\mathcal{S}_0$ is chosen with a sampling probability proportional to $1/\{1+\exp(3x_{Ai})\}$, which means that younger ones are more likely to provide telematics information due to their lower resistance to new technologies. In this case, $\delta \perp N|\mathbf{x}_\tau$.
- **Favorable selection**: Each data point assigned to $\mathcal{S}_0$ is chosen with the sampling probability proportional to $1/\{1+\exp(2N_i)\}$, which means that those with less risky behaviors are more likely to provide the telematics information. In this case, $\delta \not\perp N|\mathbf{x}_\tau$.
- **Mileage selection**: Each data point assigned to $\mathcal{S}_0$ is chosen with the sampling probability proportional to $1/\{1+\exp(x_{Mi})\}$, which means that those with lower mileage are more likely to provide the telematics information. In this case, $\delta \perp N|\mathbf{x}_\tau$.

2. After that, 80% of data points are used as a large dataset, but only with traditional features and the response variable $\{N_i, \mathbf{x}_{i\tau}\}$, which is equivalent to $\mathcal{S}_1$ in Section 2.
   For comparison, we consider the following models to estimate $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$:

   - **Naive model**: Fit a usual Poisson GLM using the data points in $\mathcal{S}_0$, which is equivalent to solving (3.3) for $\boldsymbol{\beta}$ assuming that $\omega_i = 1$ for all i.
   - **Traditional model**: Fit a usual Poisson GLM using only traditional features and the response variable $\{N_i, \mathbf{x}_{i\tau}\}$ in $\mathcal{S}_0 \cup \mathcal{S}_1$, which is equivalent to solving (3.2) for $\boldsymbol{\beta}_1$ assuming that $\boldsymbol{\beta}_2 = 0$. As such, this model does not allow the use of telematics information in the risk classification.
   - **Full model**: It uses all data points in the training set to estimate the regression coefficients, which is equivalent to solve (3.2) for $\boldsymbol{\beta}$. Therefore, it is expected to provide the best estimation performance, but may not be available in practice.
   - **Boosting model**: It uses the same estimates of $\boldsymbol{\beta}_1$ from the traditional model and computes $\hat{\eta}_i = \exp(\mathbf{x}_{i\tau}\hat{\boldsymbol{\beta}}_1)$ for each observation i in $\mathcal{S}_0$. After that, another Poisson GLM is fitted with $\mathcal{S}_0$ where the telematics information, $\mathbf{x}_{iT}$, is the only regressor and $\log \hat{\eta}_i$ is used as an offset to further estimate $\hat{\boldsymbol{\beta}}_2$ as mentioned in Ayuso *et al.* (2019). It is equivalent to solving (3.3) for $\boldsymbol{\beta}_2$ assuming that $\omega_i = 1$ for all i while $\boldsymbol{\beta}_1$ is replaced with its estimate from the traditional model.
   - **Proposed model**: It follows the estimation procedures described in Section 3, which is equivalent to solve (3.3) for $\boldsymbol{\beta}$ where ω_i is replaced by $\hat{\omega}_i(\hat{\phi})$ for all i. In this study, we use $1/v_i = \mathit{Deviance}(\text{Traditional})_i - \mathit{Deviance}(\text{Naive})_i$, where the $\mathit{Deviance}_i$ is the deviance contribution of ith individual in S_0.

3. Lastly, to incorporate the possibility that a policyholder may choose to opt for a telematics policy or not over time, 10% of data points are randomly set aside as $\mathcal{T}$ for out-of-sample validation (equivalently, the test set $\mathcal{T}$ is a representative sample of the population), where $\{N_i, \mathbf{x}_{i\tau}, \mathbf{x}_{iT}\}$ are all available.

After fitting all models, the regression estimates of these models were used to find the predicted values $\hat{N}_i = \exp(\mathbf{x}_{i\tau}\hat{\boldsymbol{\beta}}_1 + \mathbf{x}_{iT}\hat{\boldsymbol{\beta}}_2)$ for i in the out-of-sample validation set $\mathcal{T}$. Note that generation of each of the finite population, data split, regression coefficients estimation, and the out-of-sample validation are repeated $R = 1000$ times with different random seeds.

Table 1 shows the estimation results of the regression coefficients under different model specifications and sampling schemes. Here, bias, root mean square error (RMSE) and 90% confidence interval coverage (CI) of β_j are defined as follows:

$$\text{Bias}_j = \frac{1}{R}\sum_{r=1}^{R}\left(\beta_j - \hat{\beta}_j^{(r)}\right),$$

https://doi.org/10.1017/asb.2024.6 Published online by Cambridge University Press

Table 1. *Estimation performance with the simulated data (Here* ***N****,* ***T****,* ***B****,* ***F****, and* ***P*** *refer to Naive, Traditional, Boosting, Full, and Proposed models, respectively).*

	Bias					RMSE					CI				
	N	**T**	**B**	**F**	**P**	**N**	**T**	**B**	**F**	**P**	**N**	**T**	**B**	**F**	**P**
Random selection															
β_0	0.005	−0.125	−0.125	0.000	−0.019	0.190	0.142	0.142	0.066	0.080	0.902	0.376	0.376	0.893	0.858
β_{A1}	−0.018	0.004	0.004	0.002	0.004	0.851	0.300	0.300	0.295	0.338	0.902	0.877	0.877	0.881	0.869
β_{A2}	0.018	−0.006	−0.006	−0.003	−0.006	0.866	0.302	0.302	0.297	0.340	0.901	0.876	0.876	0.884	0.871
β_G	−0.001	0.000	0.000	0.000	0.000	0.055	0.019	0.019	0.018	0.021	0.905	0.879	0.879	0.890	0.871
β_M	0.001	0.000	0.000	0.000	0.000	0.017	0.005	0.005	0.005	0.007	0.907	0.911	0.911	0.916	0.906
β_T	0.000		0.048	0.000	0.041	0.027		0.053	0.009	0.052	0.901		0.313	0.887	0.458
Age selection															
β_0	0.004	−0.125	−0.125	0.000	−0.020	0.174	0.142	0.142	0.066	0.078	0.911	0.376	0.376	0.893	0.878
β_{A1}	−0.021	0.004	0.004	0.002	0.011	0.827	0.300	0.300	0.295	0.331	0.906	0.877	0.877	0.881	0.898
β_{A2}	0.027	−0.006	−0.006	−0.003	−0.013	0.896	0.302	0.302	0.297	0.339	0.904	0.876	0.876	0.884	0.895
β_G	0.000	0.000	0.000	0.000	0.000	0.052	0.019	0.019	0.018	0.021	0.907	0.879	0.879	0.890	0.889
β_M	0.001	0.000	0.000	0.000	0.000	0.016	0.005	0.005	0.005	0.007	0.902	0.911	0.911	0.916	0.900
β_T	0.001		0.050	0.000	0.038	0.026		0.054	0.009	0.051	0.901		0.234	0.887	0.518
Favorable selection															
β_0	1.478	−0.125	−0.125	0.000	0.001	1.527	0.142	0.142	0.066	0.120	0.010	0.376	0.376	0.893	0.837
β_{A1}	−0.157	0.004	0.004	0.002	0.010	1.705	0.300	0.300	0.295	0.435	0.917	0.877	0.877	0.881	0.843
β_{A2}	0.143	−0.006	−0.006	−0.003	−0.012	1.739	0.302	0.302	0.297	0.431	0.910	0.876	0.876	0.884	0.845
β_G	0.004	0.000	0.000	0.000	0.002	0.112	0.019	0.019	0.018	0.028	0.903	0.879	0.879	0.890	0.833
β_M	0.005	0.000	0.000	0.000	−0.001	0.036	0.005	0.005	0.005	0.009	0.902	0.911	0.911	0.916	0.873
β_T	0.009		0.344	0.000	0.008	0.057		0.344	0.009	0.095	0.888		0.000	0.887	0.779
Mileage selection															
β_0	−0.002	−0.125	−0.125	0.000	0.017	0.209	0.142	0.142	0.066	0.141	0.895	0.376	0.376	0.893	0.774
β_{A1}	0.012	0.004	0.004	0.002	0.046	0.924	0.300	0.300	0.295	0.601	0.891	0.877	0.877	0.881	0.768
β_{A2}	−0.014	−0.006	−0.006	−0.003	−0.047	0.937	0.302	0.302	0.297	0.608	0.884	0.876	0.876	0.884	0.780
β_G	0.002	0.000	0.000	0.000	0.000	0.058	0.019	0.019	0.018	0.039	0.894	0.879	0.879	0.890	0.756
β_M	0.001	0.000	0.000	0.000	−0.022	0.034	0.005	0.005	0.005	0.034	0.905	0.911	0.911	0.916	0.583
β_T	0.001		0.050	0.000	0.045	0.028		0.055	0.009	0.075	0.896		0.317	0.887	0.731

$$\text{RMSE}_j = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\beta_j - \hat{\beta}_j^{(r)}\right)^2},$$

$$\text{CI}_j = \frac{1}{R}\sum_{r=1}^{R}\mathbb{1}_{\{|\beta_j - \hat{\beta}_j^{(r)}| < 1.645\cdot \text{SE}(\hat{\beta})_j^{(r)}\}},$$

where $\hat{\beta}_j^{(r)}$ is the estimate of β_j at r^{th} simulation, and $\text{SE}(\hat{\beta})_j^{(r)}$ is the estimated standard error of $\hat{\beta}_j^{(r)}$.

From Table 1, it is clearly observed that if the sampling mechanism of $\mathcal{S}_0$ is purely random, then the use of the naive model is less problematic in terms of estimation performance. Although the full model shows the best performance in the estimation performance followed by the proposed model, the boosting model (and correspondingly the traditional model) suffers from the biases in $\hat{\beta}_0$ and $\hat{\beta}_T$. One can also observe that although the naive model is unbiased in the case of random selection, it is less efficient in the parameter estimation compared to both the full and proposed models as shown in the higher values of RMSE. When the sampling mechanism is age selection, it is shown that the naive model has larger biases for $\hat{\beta}_{A1}$ and $\hat{\beta}_{A2}$ compared to the full and proposed models, as these coefficients correspond to the age covariate that comes with selection biases in this scenario. On the other hand, if the sampling mechanism of $\mathcal{S}_0$ is prone to favorable selection, then the differences in estimation performance are more dramatic. Unlike the random sampling case, the naive model severely suffers from lack of fit and biases in the estimates while only the full and proposed models provide acceptable ranges of estimates as $\mathcal{S}_0$ is no longer a representative sample of the finite population. Lastly, in the case of mileage selection, there is no significant improvement in estimation performance of the proposed model compared to the naive model, but the insight from results is similar with the age selection. Note that the values of Bias, RMSE, and CI of the traditional and full models across all four sampling methods are identical, which is natural as both models do not depend on the sample split between $\mathcal{S}_0$ and $\mathcal{S}_1$ for estimation of the regression coefficients. In the case of boosting model, it also shows identical values of Bias, RMSE, and CI for all the traditional covariates as it shares the estimated coefficients with the traditional model by definition.

Note that mileage can appear in both traditional and telematics datasets as self-perceived mileage and actual mileage, respectively. If the actual mileage is used for the selection, the sampling scheme with mileage selection becomes non-ignorable. While it could be meaningful to consider the non-ignorable missing mechanism in the UBI context (choosing a UBI policy based on telematics variables) as mentioned in Boucher *et al.* (2013), handling a non-ignorable missing pattern requires to jointly model δ and $\mathbf{x}_T$ (Heckman, 1976; Glynn *et al.*, 2013) that comes with much more distributional assumptions and restrictions. In this regard, we delegate this issue as a future research topic and refrain from further discussing this issue in the current paper.

Note that the efficiency gain in the estimation of $\boldsymbol{\beta}_2 = \beta_T$ using the proposed model is no better than the naive model, unlike in the cases of $\boldsymbol{\beta}_1 = (\beta_0, \beta_{A1}, \beta_{A2}, \beta_G, \beta_M)$. It is reasonable since there is no information to borrow from $\mathcal{S}_1$ to better estimate $\boldsymbol{\beta}_2$ in the proposed model.

After assessing the estimation performance of each model, we use the out-of-sample validation set $\mathcal{T}_r$ for each $r = 1, \ldots, R$ to compare their predictive performance. In the out-of-sample validation, we use prediction RMSE (pRMSE) and the Poisson deviance statistic (DEV) defined as follows:

$$\begin{aligned}
\text{Avg_pRMSE}^{(k)} &= \frac{1}{R}\sum_{r=1}^{R}\text{pRMSE}^{(r;k)},\\
\text{pRMSE}^{(r;k)} &= \sqrt{\frac{1}{|\mathcal{T}_r|}\sum_{i\in\mathcal{T}_r}\left(N_i^{(r)} - \hat{N}_i^{(r;k)}\right)^2},\\
\text{Avg_DEV}^{(k)} &= \frac{1}{R}\sum_{r=1}^{R}\text{DEV}^{(r;k)}\\
\text{DEV}^{(r;k)} &= \frac{2}{|\mathcal{T}_r|}\sum_{i\in\mathcal{T}_r}\left[N_i^{(r)}\log\left(N_i^{(r)}/\hat{N}_i^{(r;k)}\right) + \left(N_i^{(r)} - \hat{N}_i^{(r;k)}\right)\right],
\end{aligned} \tag{4.1}$$

https://doi.org/10.1017/asb.2024.6 Published online by Cambridge University Press

Table 2. *Out-of-sample validation performance with the simulated data.*

	Naive	Traditional	Boosting	Full	Proposed
Random selection					
Avg_pRMSE	0.38192	0.38998	0.38216	0.38176	0.38187
Avg_DEV	0.55590	0.59170	0.55737	0.55534	0.55573
Age selection					
Avg_pRMSE	0.38191	0.38998	0.38215	0.38176	0.38187
Avg_DEV	0.55591	0.59170	0.55735	0.55534	0.55572
Favorable selection					
Avg_pRMSE	0.40327	0.38998	0.38596	0.38176	0.38204
Avg_DEV	0.75695	0.59170	0.57256	0.55534	0.55635
Mileage selection					
Avg_pRMSE	0.38211	0.38998	0.38215	0.38176	0.38219
Avg_DEV	0.55641	0.59170	0.55736	0.55534	0.55671

where $|\mathcal{T}_r|$ is the number of observations in $\mathcal{T}_r$ and the predicted value $\hat{N}_i^{(r;k)}$ is generated in model k with r^{th} simulation sample. Table 2 presents the out-of-sample validation performance of the aforementioned models. Again, the values of Avg_pRMSE and Avg_DEV of the traditional and full models across all four sampling methods are identical as the estimated regression coefficients, which are used for the prediction, are identical across all sampling methods. As in Table 1, the use of naive and boosting models is more vulnerable when the availability of telematics information is prone to favorable selection. It is also shown that the predictive performance of the traditional model is generally inferior to the other models, since it completely ignores the impacts of the available telematics information. Lastly, it is shown that the proposed model shows satisfactory prediction performance comparable to that of the full model (ideal yet not available in practice) in all scenarios for the missing mechanism.

5. Data analysis

5.1. Data description

To assess the validity and applicability of the proposed method under a more realistic environment than the simulation study with possible sampling biases, we use a synthetic dataset from the study of So *et al.* (2021) that includes traditional characteristics, telematics characteristics, and the response variable. As mentioned in Section 1, it has been difficult for researchers to access a dataset on insurance claims with telematics features due to privacy concerns and proprietary issues of insurers. In this regard, So *et al.* (2021) effectively emulated a synthetic dataset that shares remarkably similar statistics with the original dataset yet still preserves the privacy of the observations from the original source. Due to scarcity of a realized data split for $\mathcal{S}_0$, $\mathcal{S}_1$, and $\mathcal{T}$ that are simultaneously obtained from an actual insurance poftfolio, here we assumed that the synthetic dataset of So *et al.* (2021) is the finite population including both the traditional and telematics features while the data splits followed the sampling schemes of Section 4. Note that our purpose is not to detect selection biases from an actual insurance portfolio, but to quantify impacts of the proposed method under potential selection biases. Although the available features in the dataset are already summarized in tabular format compared to the raw data directly obtained from the telematics device, they are still high dimensional. For example, one of the "traditional" features is `Region`, which is a categorical variable with 55 categories.

However, the proposed data integration approach is based on estimating equations and GLMs so that it lacks the ability to handle high dimensionality on its own, unlike neural network models or tree-based

https://doi.org/10.1017/asb.2024.6 Published online by Cambridge University Press

Table 3. *Variable names and descriptions of the preprocessed dataset.*

Type	Variable	Description
Traditional	`Duration`	Duration of the insurance coverage of a given policy, in days
	`Insured.age`	Age of insured driver, in years
	`Insured.sex`	Sex of insured driver (Male/Female)
	`Car.age`	Age of vehicle, in years
	`Marital`	Marital status (Single/Married)
	`Car.use`	Use of vehicle: Private, Commute, Farmer, Commercial
	`Credit.score`	Credit score of insured driver
	`Region`	Type of region where driver lives: rural, urban
	`Annual.miles.drive`	Annual miles expected to be driven declared by driver
	`Years.noclaims`	Number of years without any claims
	`TerritoryEmb`	Embedded value from the territorial location of vehicle
Telematics	`Annual.pct.driven`	Annualized percentage of time on the road
	`Pct.drive.xxx`	Percent of driving day xxx of the week: mon/tue/. . ./sun
	`Pct.drive.rush.am`	Percent of driving during am rush hours
	`Pct.drive.rush.pm`	Percent of driving during pm rush hours
	`Avgdays.week`	Mean number of days used per week
	`Accel.06miles`	Number of sudden acceleration 6 mph/s per 1000 miles
	`Brake.06miles`	Number of sudden brakes 6 mph/s per 1000 miles
	`Acbr.others`	Total number of sudden acceleration and brakes 8/9/. . ./14 mph/s per 1000 miles
	`Left.turns`	Number of left turn per 1000 miles with intensity greater than equal to 8
	`Right.turns`	Number of right turn per 1000miles greater than equal to 8
Response	`NB_Claim`	Number of observed claims

models. In this regard, some of the available features were preprocessed. Due to the high dimension of the dataset and the complexity of defining some of its features, the territorial embedding and principal component analysis (PCA) were utilized to clean up the dataset. After data preprocessing, we retained the following variables for our analysis as described in Table 3. For details of data preprocessing, see Jeong (2022).

5.2. Estimation and prediction results

Unlike the simulation study, it is hardly possible to believe that the actual observations in the synthetic dataset follow the specified Poisson GLM. In this regard, here we replicate the empirical distribution of the preprocessed dataset (which is our finite population here) by generating bootstrap samples to ensure each observation has the same empirical distribution as the finite population. More specifically, we take bootstrap samples $\mathcal{S}_0$ and $\mathcal{S}_1$ of sizes 100,000 and 800,000, respectively, in each of the sampling schemes listed in Section 4. Subsequently, a bootstrap sample $\mathcal{T}$ of size 100,000 is taken at random for out-of-sample validation. After that, we repeat the process of fitting and testing these five models as in Section 4 for $R = 500$ times to compare the estimation and predictive performance under each sampling scheme.

To assess the in-sample estimation performance, we compare the estimated regression coefficients of each method and sampling scheme with the estimated regression coefficients obtained from the finite population. More specifically, bias, root mean squared error (RMSE), and 90% confidence interval coverage (CI) of the regression coefficients are defined as follows:

https://doi.org/10.1017/asb.2024.6 Published online by Cambridge University Press

$$\text{Bias}_j = \frac{1}{R}\sum_{r=1}^{R}\left(\tilde{\beta} - \hat{\beta}_j^{(r)}\right),$$

$$\text{RMSE}_j = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\tilde{\beta} - \hat{\beta}_j^{(r)}\right)^2},$$

$$\text{CI}_j = \frac{1}{R}\sum_{r=1}^{R}\mathbb{1}_{\left\{|\tilde{\beta}_j - \hat{\beta}_j^{(r)}| < 1.645\cdot \text{SE}(\hat{\beta})_j^{(r)}\right\}},$$

where $\tilde{\beta}_j$ and $\hat{\beta}_j^{(r)}$ are estimates of β_j using the finite population and with r^{th} bootstrap sample, respectively. $\text{SE}(\hat{\beta})_j^{(r)}$ is the estimated standard error of $\hat{\beta}_j^{(r)}$. Note that in our case, we prefer a method with biases closer to 0, smaller RMSEs, and/or CIs closer to the theoretical benchmark, 90%.

Table S1 shows the estimation results of the regression coefficients under different model specifications and sampling schemes of the bootstrap samples from the prerocessed synthetic data. Note that the estimated coefficients from the traditional model were omitted as they are only available for the traditional features and identical to those from the boosting model. Implications of the estimation results with the actual data are as follows.

- In the case of random selection, only the boosting model suffers from the biases of the regression coefficients, and there are no big differences in the estimation performance between the naive and proposed models. It implies that as long as the sampling mechanism of $\mathcal{S}_0$ (a small dataset with both traditional and telematics features) from the finite population is purely random, it is okay to ignore $\mathcal{S}_1$ (a large dataset only with traditional features) and analyze $\mathcal{S}_0$ for ratemaking purposes.
- In the case of age selection and mileage selection, the naive model is more biased in the estimation of the traditional covariates (especially the intercept term) compared to the proposed model. It implies that if the observability of the telematics features depends on the traditional features, then the proposed approach might be helpful in better understanding the underlying impacts of the covariates on the claim experience.
- Lastly, in the case of favorable selection, the proposed model is no more unbiased, but the naive model is still more biased in the estimation of the regression coefficients. Therefore, if accessibility to telematics features is affected by favorable selection, it is recommended to integrate two data sources to handle the missingness and selection biases of the telematics features.

Such differences are also visualized in Figures S1, S2, S3, and S4 where a model with biases closer to 0, smaller RMSEs, and/or CIs closer to 90% receives the higher rank for each covariate. It is consistently observed that, in the case of either age or favorable selection, the proposed model is the second best, following the full model that is unattainable in practice.

In addition to the estimation performance, the out-of-sample validation performance is assessed using

$$\text{Avg\_pRMSE}^{(k)} = \frac{1}{R}\sum_{r=1}^{R}\text{pRMSE}^{(r;k)},$$

$$\text{Prop\_pRMSE}^{(k)} = \frac{1}{R}\sum_{r=1}^{R}\mathbb{I}\left(\text{pRMSE}^{(r;k)} > \text{pRMSE}^{(r;proposed)}\right),$$

$$\text{Avg\_DEV}^{(k)} = \frac{1}{R}\sum_{r=1}^{R}\text{DEV}^{(r;k)}$$

$$\text{Prop\_DEV}^{(k)} = \frac{1}{R}\sum_{r=1}^{R}\mathbb{I}\left(\text{DEV}^{(r;k)} > \text{DEV}^{(r;proposed)}\right),$$

https://doi.org/10.1017/asb.2024.6 Published online by Cambridge University Press

Table 4. *Out-of-sample validation performance with bootstrapping from the actual data.*

	Naive	Traditional	Boosting	Full	Proposed
Random selection					
Avg_pRMSE	0.21187	0.21621	0.21204	0.21181	0.21184
Prop_pRMSE	0.653	1.000	0.998	0.309	–
Avg_DEV	24.01372	26.74222	24.20771	23.99236	24.01203
Prop_DEV	0.526	1.000	1.000	0.020	–
Age selection					
Avg_pRMSE	0.21194	0.21619	0.21214	0.21179	0.21183
Prop_pRMSE	0.801	1.000	0.998	0.229	–
Avg_DEV	24.03209	26.73960	24.20694	23.98884	24.00955
Prop_DEV	0.917	1.000	1.000	0.012	–
Favorable selection					
Avg_pRMSE	0.21402	0.21620	0.21410	0.21180	0.21187
Prop_pRMSE	1.000	1.000	1.000	0.182	–
Avg_DEV	25.48580	26.74150	25.49272	23.99129	24.02208
Prop_DEV	1.000	1.000	1.000	0.004	–
Mileage selection					
Avg_pRMSE	0.21188	0.21612	0.21200	0.21171	0.21176
Prop_pRMSE	0.912	1.000	1.000	0.184	–
Prop_DEV	24.02466	26.74138	24.20250	23.99120	24.01350
Prop_DEV	0.744	1.000	1.000	0.016	–

where $\text{pRMSE}^{(r;k)}$ and $\text{DEV}^{(r;k)}$ are defined in (4.1). Based on the above definition, we prefer a model with lower Avg_pRMSE, Prop_pRMSE, Avg_DEV, and/or Prop_DEV.

Table 4 shows that the proposed model is the only model comparable to the full model in terms of pRMSE and DEV on average, especially when the observability of telematics features is prone to favorable selection. It is also observed that the naive, traditional, and boosting models do not outperform the proposed model in most bootstrap samples, as shown in the values of Prop_pRMSE and Prop_DEV, regardless of the selection scheme. Therefore, the proposed approach is a reasonable alternative in the absence of a finite population with both traditional and telematics features.

Figure S5 further highlights the distributions of proportional improvements in pRMSE and DEV using the proposed model compared to the naive model, where the proportional improvements in pRMSE or DEV with r^{th} bootstrap sample are defined as $100\left(1-\frac{\text{pRMSE}^{(r;proposed)}}{\text{pRMSE}^{(r;naive)}}\right)$ and $100\left(1-\frac{\text{DEV}^{(r;proposed)}}{\text{DEV}^{(r;naive)}}\right)$, respectively. While proportional improvements are shown to be close to symmetric and almost centered on 0 with positive averages for random, age or mileage selection, they are clearly positive with favorable selection, which also supports the usefulness of the proposed method on the existence of favorable selection in the provision of telematics features.

6. Concluding remarks

The scarcity of observations with telematics features in driver risk classification for auto insurance has been problematic, which may be attributed to either privacy concerns or favorable selection when compared to traditional feature data points. To address this issue, we proposed a data integration approach that uses calibration weights for UBI with multiple sources of insurance claims data. Our results

https://doi.org/10.1017/asb.2024.6 Published online by Cambridge University Press

demonstrate that this framework can effectively integrate traditional and telematics data, while also managing potential favorable selection problems. This conclusion is supported by a simulation study and empirical analysis using a synthetic telematics dataset as it turns out that the proposed approach could achieve satisfactory performance both in the in-sample estimation and in the out-of-sample prediction, compared to the existing benchmarks for automobile insurance ratemaking practices. Thus, the proposed approach has a potential to improve risk classification in auto insurance and assist insurers in making informed decisions.

The possible extension of this article is twofold. First, the proposed data integration approach relies on the assumption in (3.5) so it might not work well if the basis function of $E\{U(\boldsymbol{\beta}; \mathbf{x}_i, n_i) \mid \mathbf{x}_{i\tau}, n_i\}$ is not correctly specified. To address such a problem, one can implement a doubly robust calibration approach that only requires either the basis function of the outcome variable or the propensity score to be correctly specified. Second, the proposed approach can be extended to data integration for mixed-effects models where a policyholder is observed over a period of time, so that the proposed framework can also consider random effects for experience ratemaking, as well as the fixed effects.

Supplementary material. To view supplementary material for this article, please visit https://doi.org/10.1017/asb.2024.6

References

Agresti, A. (2003) *Categorical Data Analysis*. Hoboken, NJ: John Wiley & Sons, Inc.

Ayuso, M., Guillen, M. and Nielsen, J.P. (2019) Improving automobile insurance ratemaking using telematics: Incorporating mileage and driver behaviour data. *Transportation*, **46**(3), 735–752.

Ayuso, M., Guillén, M. and Pérez-Marn, A.M. (2014) Time and distance to first accident and driving patterns of young drivers with pay-as-you-drive insurance. *Accident Analysis and Prevention*, **73**, 125–131.

Ayuso, M., Guillen, M. and Pérez-Marn, A.M. (2016) Telematics and gender discrimination: some usage-based evidence on whether men's risk of accidents differs from women's. *Risks*, **4**(2), 10.

Baecke, P. and Bocca, L. (2017) The value of vehicle telematics data in insurance risk selection processes. *Decision Support Systems*, **98**, 69–79.

Boucher, J.-P., Pérez-Marn, A.M. and Santolino, M. (2013) Pay-as-you-drive insurance: The effect of the kilometers on the risk of accident. In *Anales del Instituto de Actuarios Españoles*, vol. 19, pp. 135–154. Instituto de Actuarios Españoles Madrid.

Cather, D.A. (2020) Reconsidering insurance discrimination and adverse selection in an era of data analytics. *The Geneva Papers on Risk and Insurance-Issues and Practice*, **45**, 426–456.

Cohen, A. and Siegelman, P. (2010) Testing for adverse selection in insurance markets. *Journal of Risk and Insurance*, **77**(1), 39–84.

Denuit, M., Guillen, M. and Trufin, J. (2019) Multivariate credibility modelling for usage-based motor insurance pricing with behavioural data. *Annals of Actuarial Science*, **13**(2), 378–399.

Derikx, S., De Reuver, M. and Kroesen, M. (2016) Can privacy concerns for insurance of connected cars be compensated? *Electronic Markets*, **26**, 73–81.

Dewri, R., Annadata, P., Eltarjaman, W. and Thurimella, R. (2013) Inferring trip destinations from driving habits data. *Proceedings of the 12th ACM workshop on Workshop on Privacy in the Electronic Society*, pp. 267–272.

Duri, S., Gruteser, M., Liu, X., Moskowitz, P., Perez, R., Singh, M. and Tang, J.-M. (2002) Framework for security and privacy in automotive telematics. *Proceedings of the 2nd International Workshop on Mobile Commerce*, pp. 25–32.

Duval, F., Boucher, J.-P. and Pigeon, M. (2023) Enhancing claim classification with feature extraction from anomaly-detection-derived routine and peculiarity profiles. *Journal of Risk and Insurance*, **90**(2), 421–458.

Eling, M. and Kraft, M. (2020) The impact of telematics on the insurability of risks. *Journal of Risk Finance*, **21**(2), 77–109.

Gao, G., Meng, S. and Wüthrich, M.V. (2019) Claims frequency modeling using telematics car driving data. *Scandinavian Actuarial Journal*, **2019**(2),143–162.

Gao, G., Wang, H. and Wüthrich, M.V. (2022) Boosting Poisson regression models with telematics car driving data. *Machine Learning*, **111**(1), 243–272.

Glynn, R.J., Laird, N.M. and Rubin, D.B. (2013) Selection modeling versus mixture modeling with nonignorable nonresponse. In *Drawing Inferences from Self-Selected Samples*, pp. 115–142. Routledge.

Guillen, M., Nielsen, J.P. and Pérez-Marn, A.M. (2021) Near-miss telematics in motor insurance. *Journal of Risk and Insurance*, **88**(3), 569–589.

Heckman, J.J. (1976) The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. In *Annals of Economic and Social Measurement*, vol. 5, number 4, pp. 475–492. NBER.

Holzapfel, J., Peter, R. and Richter, A. (2023) Mitigating moral hazard with usage-based insurance. *Journal of Risk and Insurance*, 1–27. https://doi.org/10.1111/jori.12433.

Husnjak, S., Peraković, D., Forenbacher, I. and Mumdziev, M. (2015) Telematics system in usage based motor insurance. *Procedia Engineering*, **100**, 816–825.
Imai, K. and Ratkovic, M. (2014) Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **76**, 243–263.
Jeong, H. (2022) Dimension reduction techniques for summarized telematics data. *Journal of Risk Management*, **33**(4), 1–25.
Klugman, S.A., Panjer, H.H. and Willmot, G.E. (2012) *Loss Models: From Data to Decisions*, vol. 715. Hoboken, NJ: John Wiley & Sons, Inc.
Ma, Y.-L., Zhu, X., Hu, X. and Chiu, Y.-C. (2018) The use of context-sensitive insurance telematics data in auto insurance rate making. *Transportation Research Part A: Policy and Practice*, **113**, 243–258.
MarketsandMarkets (2021) URL https://www.marketsandmarkets.com/Market-Reports/usage-based-insurance-market-154621760.html.
Milanović, N., Milosavljević, M., Benković, S., Starčević, D. and Spasenić, Ž. (2020) An acceptance approach for novel technologies in car insurance. *Sustainability*, **12**(24), 10331.
NAIC (2015) URL https://content.naic.org/cipr_special_reports.htm.
Reimers, I. and Shiller, B.R. (2019) The impacts of telematics on competition and consumer behavior in insurance. *The Journal of Law and Economics*, **62**(4), 613–632.
Rothschild, M. and Stiglitz, J. (1978) Equilibrium in competitive insurance markets: An essay on the economics of imperfect information. *The Quarterly Journal of Economics*, **90**(4),629–649.
Rubin, D.B. (1976) Inference and missing data. *Biometrika*, **63**, 581–592.
Śliwiński, A. and Kuryowicz,. (2021) Usage-based insurance and its acceptance: An empirical approach. *Risk Management and Insurance Review*, **24**(1), 71–91.
So, B., Boucher, J.-P. and Valdez, E.A. (2021) Synthetic dataset generation of driver telematics. *Risks*, **9**(4), 58.
Verbelen, R., Antonio, K. and Claeskens, G. (2018) Unravelling the predictive power of telematics data in car insurance pricing. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **67**(5), 1275–1304.
Wang, H. and Kim, J.K. (2021) Information projection approach to propensity score estimation for handling selection bias under missing at random. arXiv e-prints, arXiv–2104.
Yang, S. and Kim, J.K. (2020) Statistical data integration in survey sampling: A review. *Japanese Journal of Statistics and Data Science*, **3**(2), 625–650.

Appendix

A. Proof of Proposition 1

Now, as long as (3.6) is satisfied, we can express

$$\begin{aligned}\sum_{i\in S_0}\omega_i U(\boldsymbol{\beta};\mathbf{x}_i,n_i) &= \sum_{i=1}^{M}\delta_i\omega_i U(\boldsymbol{\beta};\mathbf{x}_i,n_i)+\sum_{l=0}^{L}\left[\alpha_l\left(\sum_{i=1}^{M}b_{li}-\sum_{i=1}^{M}\delta_i\omega_i b_{li}\right)\right]\\ &= \sum_{i=1}^{M}\delta_i\omega_i U(\boldsymbol{\beta};\mathbf{x}_i,n_i)+\sum_{i=1}^{M}(1-\delta_i\omega_i)\sum_{l=0}^{L}\alpha_l b_{li}\\ &= \sum_{i=1}^{M}\left\{\delta_i U(\boldsymbol{\beta};\mathbf{x}_i,n_i)+(1-\delta_i)\sum_{l=0}^{L}\alpha_l b_{li}\right\}\\ &\quad+\sum_{i=1}^{M}\delta_i(\omega_i-1)\left\{U(\boldsymbol{\beta};\mathbf{x}_i,n_i)-\sum_{l=0}^{L}\alpha_l b_{li}\right\}\end{aligned}$$

for any $\boldsymbol{\alpha}=(\alpha_0,\alpha_1,\ldots,\alpha_L)$. Thus, for the choice of $\hat{\boldsymbol{\alpha}}$ satisfying

$$\sum_{i=1}^{M}\delta_i(\omega_i-1)\left\{U(\boldsymbol{\beta};\mathbf{x}_i,n_i)-\sum_{l=0}^{L}\hat{\alpha}_l b_{li}\right\}=0, \tag{A1}$$

we can obtain

$$\sum_{i\in S_0}\omega_i U(\boldsymbol{\beta};\mathbf{x}_i,n_i)=\sum_{i=1}^{M}\left\{\delta_i U(\boldsymbol{\beta};\mathbf{x}_i,n_i)+(1-\delta_i)\sum_{l=0}^{L}\hat{\alpha}_l b_{li}\right\}. \tag{A2}$$

Furthermore, the condition in (A1) under model (3.5) implies that $\sum_{l=0}^{L}\hat{\alpha}_l b_{li}$ is an estimator of $E\{U(\boldsymbol{\beta};\mathbf{x}_i,n_i)\mid\mathbf{x}_{i\tau},n_i\}$. Thus, we can see that (A2) shows self-efficiency in (3.4). That is, the calibration condition (3.6) on the basis functions in (3.5) is a sufficient condition for self-efficiency.

https://doi.org/10.1017/asb.2024.6 Published online by Cambridge University Press

B. Proof of Proposition 2

To show self-efficiency in (3.4),

$$\begin{aligned}(\text{RHS}) &= \sum_{i=1}^{M} \hat{U}_i + \sum_{i\in S_0} U_i - \sum_{i\in S_0} \hat{U}_i \\ &= \sum_{i\in S_0} \hat{\omega}_i \hat{U}_i + \sum_{i\in S_0} U_i - \sum_{i\in S_0} \hat{U}_i \\ &= \sum_{i\in S_0} \hat{U}_i + \frac{M_1}{M_0} \sum_{i\in S_0} g_i(\hat{\phi}) \hat{U}_i + \sum_{i\in S_0} U_i - \sum_{i\in S_0} \hat{U}_i \\ &= \frac{M_1}{M_0} \sum_{i\in S_0} g_i(\hat{\phi}) U_i + \sum_{i\in S_0} U_i = (\text{LHS}),\end{aligned}$$

where the second equality follows from (3.6) and the fourth equality follows from (3.9).

C. Standard error estimation

The standard errors of the estimates can be estimated using the standard linearization method. Note that $\boldsymbol{\beta}$ is the parameter of interest, and $\boldsymbol{\phi}$ is the nuisance parameter that is used to estimate the parameter of interest $\boldsymbol{\beta}$. To estimate the variance of $\hat{\boldsymbol{\beta}}$, we also need to estimate the variance of $\hat{\boldsymbol{\phi}}$ simultaneously. Thus, we can construct two estimating functions for two parameters as follows.

$$\begin{aligned}\hat{U}_1(\boldsymbol{\phi}) &= \sum_{i\in\mathcal{S}} \{\delta_i \omega_i(\boldsymbol{\phi}) - 1\} \mathbf{b}_i, \\ \hat{U}_2(\boldsymbol{\phi}, \boldsymbol{\beta}) &= \sum_{i\in\mathcal{S}} \delta_i \hat{\omega}_i(\boldsymbol{\phi}) U(\boldsymbol{\beta}; \mathbf{x}_i, n_i),\end{aligned}$$

where $\mathbf{b}_i = (1, b_{1i}, \cdots, b_{Li})'$ and

$$\omega_i(\boldsymbol{\phi}) = 1 + \frac{M_1}{M_0} \exp\{\phi_0 + \phi_1 b_{1i} + \cdots + \phi_L b_{Li}\}.$$

The final estimator $\hat{\boldsymbol{\beta}}$ is the solution to the joint estimating equations:

$$\hat{U}_1(\boldsymbol{\phi}) = 0 \quad \text{and} \quad \hat{U}_2(\boldsymbol{\phi}, \boldsymbol{\beta}) = 0.$$

We can treat $\boldsymbol{\theta}' = (\boldsymbol{\phi}', \boldsymbol{\beta}')$ and define

$$\hat{U}(\boldsymbol{\theta}) = \begin{pmatrix} \hat{U}_1(\boldsymbol{\phi}) \\ \hat{U}_2(\boldsymbol{\phi}, \boldsymbol{\beta}) \end{pmatrix}.$$

The variance estimation for $\hat{\boldsymbol{\theta}}$ can be implemented using the Sandwich formula. That is, $V(\hat{\boldsymbol{\theta}}) = \tau^{-1} V(\hat{U}) \tau^{-1'}$ where $\tau = E\left\{\frac{\partial}{\partial \boldsymbol{\theta}'} \hat{U}(\boldsymbol{\theta})\right\}$.

One can use an empirical estimate of $V(\hat{\boldsymbol{\theta}})$ as follows:

$$\tilde{\tau} = \frac{\partial}{\partial \theta'} \hat{U}(\theta)\Big|_{\theta=\hat{\theta}} \qquad \text{and } \tilde{V}(\hat{U}) = \sum_{i=1}^{M} (\tilde{U}_i - \overline{\tilde{U}}_i)(\tilde{U}_i - \overline{\tilde{U}}_i)'$$

as a proxy of τ and $V(\hat{U})$, respectively, where $\hat{\theta}' = (\hat{\phi}', \hat{\boldsymbol{\beta}}')$ is the solution of the joint estimating equation and

$$\tilde{U}_i = \begin{pmatrix} \left\{\delta_i \omega_i(\hat{\phi}) - 1\right\} \mathbf{b}_i \\ \delta_i \hat{\omega}_i(\hat{\phi}) U(\hat{\boldsymbol{\beta}}; \mathbf{x}_i, y_i) \end{pmatrix}, \quad \overline{\tilde{U}}_i = \frac{1}{M} \sum_{i=1}^{M} \tilde{U}_i.$$

https://doi.org/10.1017/asb.2024.6 Published online by Cambridge University Press